Model Evaluation Workgroup

Technical Memorandum 1

Model Evaluation Metrics

Prepared By:

Limno-Tech, Inc. for the Fox River Group of Companies

and

Wisconsin Department of Natural Resources

March 13, 1998

TABLE OF CONTENTS

1.0 SUMMARY

This technical memorandum is provided in partial fulfillment of the Memorandum of Agreement ("Agreement") between the State of Wisconsin and seven paper companies ("Companies"), dated January 31, 1997.

Model evaluations will be undertaken according to the procedures discussed in the "Workplan to Evaluate the Fate and Transport Models for the Fox River and Green Bay" ("Workplan"), which was submitted by Limno-Tech, Inc. ("LTI") on behalf of the Companies to the Wisconsin Department of Natural Resources ("WDNR") on September 19, 1997. The Workplan was conditionally approved by WDNR on September 26, 1997. This technical memorandum is the product of Task 1, entitled "Development and Prioritization of Model Evaluation Metrics", as specified in the approved Workplan.

The evaluation metrics to be employed are summarized below, in order of priority (by major grouping). The individual metrics are discussed in detail in the body of this memorandum.

- Evaluation of the mathematical representation of the natural system:
  - review of conceptual basis
  - evaluate appropriateness of models for spatial and temporal context and data availability
- Short-term and long-term hindcast simulation metrics:
  - time series comparisons (predicted vs. observed) for the simulation period for:
    - water-column TSS and PCBs
    - spatial (vertical and horizontal) PCB distributions in sediment
    - fish PCB body burdens
  - point-in-time and cumulative performance diagnostic comparisons (predicted vs. observed) for the simulation period for:
    - end-of-period sediment PCB concentrations, including spatial distribution
    - end-of-period PCB mass balance
    - sediment bed elevation changes
    - net burial rates
  - comparison of distributions (predicted vs. observed) for the simulation period, for:

  - water column TSS and PCBs
  - sediment PCB concentrations
  - fish PCB body burdens
- event and non-event concentration and (where possible) flux comparisons, predicted vs. observed or computed, from the short-term simulation, for:
  - water-column TSS and PCBs
  - estimate uncertainty bounds of predictions

- Forecast simulation metrics:
  - comparative analysis of estimated uncertainty bounds relative to forecasts for different remedial strategies within any single suite of models, as well as any alternative model formulations

The ranking shown above is based on the following principles:

- that the mathematical representation of physical, chemical, and biological processes in a model (embodied by model assumptions) must represent the essential characteristics of the natural system in a manner that can be supported by comparisons between model predictions and observations
- short-term simulations for data-rich periods provide the best means to develop model procedures to assign model parameters for data-poor, hindcast periods, and long-term forecasts (i.e. simulation techniques) since forcing functions are well-defined
- long-term, retrospective hindcast simulations provide the best means to confirm the performance of model simulation techniques developed from short-term simulations and establish the long-term predictive accuracy of a model
- that accuracy (as evaluated by a suite of quantitative evaluation methods) is more important than precision (as evaluated by model uncertainty bounds).

This list is intended as an ordered yet flexible set of procedures. The prioritization of the metrics provides a guide to those tests that are potentially useful, and to the order in which to employ these tests.

This memorandum begins with an overview of the process of model development, short-term simulations (calibration), and long-term simulation (hindcasts and forecasts). This discussion focuses on the relative roles and importance of the mathematical representation of the natural system to be modeled, short-term simulations, hindcast simulations, and forecast evaluations in producing accurate and precise predictions.

The memorandum then discusses the proposed metrics in detail, indicating the usefulness and limitations of each. The memorandum concludes by establishing the priority order for the use of the metrics, which will be used when necessary to resolve conflicts between evaluations based on individual metrics.

2.0 INTRODUCTION

2.1 DEFINITION OF EXISTING AND ALTERNATIVE MODELS

To complete the model evaluation process as described in the Agreement, the existing Fox River/Green Bay models must be defined. For this evaluation process the existing models, as well as any proposed alternatives, are defined as a suite of models that will have:

1. consistent spatial and temporal domains;
2. consistent representations of state variables for particles and contaminants that allow completion of short-term and long-term, retrospective simulations; and
3. consistent use of the most complete evaluation of external forcing functions, boundary conditions, and initial conditions available.

A key aspect of the model evaluation process is to evaluate the effect of alternative representations of environmental processes on model performance. Since the intent of this evaluation process is to isolate and evaluate the effect of proposed alternative process representations, all other aspects of the models (forcing functions, boundary condition, initial conditions, state variables, etc.) must be consistent between each alternative and the existing models. In this way, the construction of alternative models will be focused on investigating the effect of specific environmental process representations such as particle deposition, erosion, and accumulation/burial. Therefore, all model alternatives advanced for evaluation by any party must share the above general features with the existing models to be considered for inclusion in State of Wisconsin led Superfund NRDA and RI/FS efforts.

2.2 OVERVIEW OF MODEL EVALUATION METRICS

The objective of this technical memorandum is to specify a set of methodologies that will:

- facilitate evaluation of the predictive capabilities of the existing models
- suggest enhancements and/or improvements to existing models, if indicated
- facilitate evaluation of alternative models, and
- facilitate selection of the most appropriate model among competing alternatives.

The process of model development and validation requires quantitative as well as some qualitative analysis. Ideally, this process proceeds sequentially and systematically, as follows:

Conceptual Model: The first step is the development and validation of a conceptual model. The model should be based on scientifically accepted theory and embody any important cause-and-effect relationships demonstrated by that theory. The mathematical representation of the physical, chemical, and biological characteristics of the modeled system should be consistent with the observed behavior of the real system and appropriate for the planned application, including spatial,

temporal, and kinetic contexts. By building on a foundation of accepted theoretical relationships, the extent to which a model can be demonstrated to appropriately represent the behavior of the natural system can be more readily established. Although qualitative, this foundation then promotes the acceptance of a model as a useful tool for making meaningful predictions of natural system behavior.

Short-term Simulation: The second step is the development of a short-term simulation. The intent of a short-term simulation is to take advantage of data-rich time periods where data quality/certainty is greatest in order to: 1) determine appropriate ranges of values for each model parameter and 2) develop generalized methods to assign model parameter values for any given environmental condition as a function of independent observations (such as flow, wind speed, temperature, etc.). For conditions within an otherwise data-rich period where no data to assign model parameters as a function of independent observations exist (such as extreme flow events), well accepted scientific principles and empirical information regarding the physical processes that control system behavior are used to fill in these gaps. Once model predictions are generated, a series of graphical and statistical analyses that are the same metrics as those subsequently used to characterize model performance for the retrospective hindcast simulation, are performed to characterize the accuracy of model performance during the short-term period. Uncertainty analysis of short-term simulation results is used to quantify the precision of model results and estimate the magnitude of model parameter uncertainty.

Long-Term, Retrospective Hindcast Simulation: The third step is the development and analysis of a long-term, retrospective hindcast to evaluate the model□s ability to simulate long-term historical trends. The intent of a long-term hindcast is to: 1) provide a check of the methods used to assign models parameter values over long timeframes as well as conditions outside of the short-term calibration, and 2) establish the predictive capabilities of the model by comparative analysis of hindcast results to short-term simulation results for the same period. When used in this manner, a hindcast is a powerful tool that provides insight into the predictive powers of the models and help quantify the abilities of models to forecast future environmental conditions. This step is crucial because a failure to accurately track the historical record may indicate potential errors in the ability to predict future trends. Once model predictions are generated, a series of graphical and statistical analyses that are the same metrics as those used to characterize model performance for the short-term simulation, are performed to characterize model performance during the hindcast period. Comparative analysis of the short-term and hindcast simulation results are used to characterize and quantify the estimated long-term accuracy of model prediction. Uncertainty analysis of the hindcast results is used to quantify the precision of model results and estimate the magnitude of model forcing function uncertainty

It should be noted that, as a consequence of processes that may not express observable effects in the system over shorter time horizons, a short-term simulation alone may not necessarily provide a sufficient means to establish the full predictive accuracy of a model. A long-term hindcast simulation provides a means to evaluate these aspects of model performance.

Future Forecasting: A fourth step is the development and analysis of forecasts (presumably for long timeframes) for future conditions. The intent of long-term forecast simulations is to predict system response(s) to proposed environmental management strategy alternatives. In deterministic

models, future conditions are typically represented as a "replay" of historical conditions or are otherwise synthesized based on statistical analysis of the frequency of occurrence of historical conditions. Analysis of forecast simulations results is based on comparative analysis of uncertainty bounds. Prediction uncertainty is expected to increase as the time period simulated increases; as uncertainty bounds overlap, it becomes more difficult to distinguish between the outcomes of simulation. This type of analysis can be used to define the point in time after which a model is no longer able to differentiate between the outcomes of environmental management strategy alternatives. Future forecasts also provide a means to evaluate model performance for situations where competing model formulation alternatives exist. In these situations, all other conditions being equal, the preferred model formulation will be the alternative with the smaller uncertainty bounds (i.e. greater precision).

To the greatest extent possible, model evaluations should be based on quantitative procedures to characterize the variability of external forcing functions, and internal parameters. The intent of quantifying the model evaluation process is to reduce the potential that arbitrary, unsupported, or ill-founded opinion and/or judgment will become the basis for, or a significant component of, model evaluations. However, no degree of quantification can entirely replace the need for qualitative analysis in model development. Qualitative analyses are often embodied as professional judgment and are frequently applied at each stage of model development, calibration, and forecasting. These judgments draw upon both objective and subjective evaluations, on the scientific literature, on key temporal, spatial, and causal relationships between variables that can be defined through analyses of field or laboratory observations, as well as from experience gained through simulations completed for other, physically similar, natural systems. For these reasons, the evaluation of model performance cannot be reduced to a set of simple, unchanging procedures. Nonetheless, the potential for developing arbitrary or otherwise inappropriate models is significantly reduced through application of concise, quantitative model evaluation metrics. This is the approach that has been adopted for the methodology developed in this document.

Finally, it is important to understand that the metrics discussed below are also intended as tools to assist in decisions to determine whether additional model development activities are desirable. These quantitative metrics reduce reliance on arbitrary judgments thereby permitting an objective evaluation of model performance.

2.3 THE NATURE OF QUANTITATIVE MODEL QUALITY CRITERIA

Similar to data quality objectives (DQOs) in a quality assurance project plan (QAPP), quantitative model quality criteria are intended to represent the target threshold of accuracy for model predictions. Ideally, model quality criteria would be specified at the outset of any model development effort and would be defined by the intended management uses of the model outputs. These criteria would then serve as a guide to establish what levels of field data collection/quality and model development effort are needed to permit development of a model that meets the desired model quality criteria. Where models are developed from independent, previously collected data sets or existing models are to be evaluated, the link between field data collection and model development efforts is decoupled. In these situations it may be appropriate to iteratively establish model quality criteria following comprehensive data review, consideration of the fundamental understanding of processes that control model predictions, and the intended management use of

model predictions.

The Agreement between the Companies and the State calls for the existing suite of Fox River/Green Bay models to be evaluated. To accomplish this, quantitative model quality criteria must be established. There are two possible paths by which to establish these criteria. The preferred approach is to develop criteria based on management requirements established by all parties prior to initiating the model evaluation process. This approach is preferred since acceptable limits of model performance are defined *a priori*. Although less desirable, an alternative approach is to establish model quality criteria based on the results of an initial evaluation of the existing models. This approach is inherently less desirable because it could result in open-ended model development and potentially permits endless advancement of model alternatives on the basis of improved relative performance (i.e. a model is never good enough because a "better" one can always be constructed).

To evaluate the existing models, the following model quality criteria are proposed: mean predicted concentrations for TSS and PCBs should be within +/- 30% of observed values for water and sediment and within +/- a factor of 3 (+/- □ order of magnitude) for short-term simulations and +/- 50% for water and sediments and a factor of 5 for fish for long-term simulations. These criteria are based on the following:

- The model development goal of the Green Bay Mass Balance Study from which the existing models were developed (at least as expressed for the river models) was to achieve agreement of +/- 30% between model predictions and observations for water and sediment and +/- □ order of magnitude for fish; and

- The performance of the existing models, as determined by previous WDNR model evaluation efforts, established agreement of +/- 30% between predictions and observations based on a blind post-audit short-term simulation of the Fox River models.

- As a result of uncertainty in forcing functions as well as model parameters, model performance for long-term simulation is expected to be somewhat less than performance for short-term simulations.

The Workplan also provides a mechanism for alternative models to be constructed and evaluated as part of the Agreement. The intent of this mechanism is to permit exploration of whether alternative representations of environmental processes yield models that are both quantifiably better (greater predictive accuracy and precision) and lead to significantly different management conclusions. To facilitate the evaluation of model alternatives that may be advanced in later stages of the evaluation process, the performance of the existing Fox River/Green Bay models is proposed as the quantitative model quality criterion that defines the minimum acceptable threshold of model performance for comparisons between model alternatives.

3.0 METRICS AND QUALITY CRITERIA

A list of metrics and quality criteria to evaluate fate and transport for PCBs in the Fox River and Green Bay can be separated into four categories: 1) mathematical representation of the natural system; 2) short-term simulation metrics; 3) long-term hindcast metrics; and 4) forecast metrics. Each category of metrics is important for judging the long-term predictive ability of these models, fate and exposure pathways, ability to adequately simulate contaminant fate, and ultimately their utility as tools to support management decision making. The prioritization of these metrics is presented in Section 4.0.

3.1 MATHEMATICAL REPRESENTATION OF THE NATURAL SYSTEM

The purpose of this metric is to establish whether the mathematical representations of the physical, chemical, and biological processes and characteristics of a model appropriately represent the behavior and characteristics of the natural system to be simulated. An important aspect of a model's theoretical soundness is appropriateness to the temporal, spatial, and kinetic context to which it is to be applied. The system to which the suite of Fox River/Green Bay models is applied consists of the following physical features:

- the Lower Fox River, including water column and sediments, from Lake Winnebago to the river mouth at Green Bay; and
- Green Bay, including water column and sediments from its head at Green Bay to it boundary with Lake Michigan.

The following transport and fate processes are represented in the models:

- erosion and deposition of particles and associated PCBs between the water column and sediment bed;
- burial of PCBs within the sediment bed;
- transformations between particle/sorbent compartments (sorbent dynamics);
- partitioning of PCBs between water and particulate phases, including dissolved organic carbon;
- advective and dispersive transport, in the water column, of suspended particles as well as adsorbed and dissolved PCBs;
- diffusive transport in sediment pore water (dissolved sorbents and dissolved/bound PCBs);
- air-water exchange of PCBs (wet/dry deposition, volatilization/gas absorption); and
- PCB accumulation (direct uptake, trophic transfer, depuration, etc.) in biota.

The suite of models will be used to forecast contaminant levels in water, sediments, and fish over a scale of decades. The models will be used to assist in:

- PCB exposure pathway determination and analysis of restoration alternatives for the Natural Resource Damage Assessment; and
- remedial planning activities, including the no action alternative and other specified remedial alternatives.

Evaluations of the mathematical representations of the models shall address their appropriateness for this time frame and these applications, and include the physical features and fate and transport processes listed above.

A model that is appropriate for the calibration period and/or hindcasting period may not be appropriate for the forecast period. This could be because future site conditions are expected to be different from conditions that prevailed in the past, or because of proposed management alternatives that were not attempted in the past. The evaluation of the model's mathematical representation of the natural system shall include its appropriateness for future, as well as past and present, conditions.

Whenever mathematical representation issues are raised, the model evaluation workgroup shall discuss implications, establish importance, and indicate possible alternatives.

3.2 SHORT-TERM SIMULATION METRICS

These include quantitative (including statistical) and qualitative comparisons of model predictions and observations, for the short-term simulation time period of 1989 through 1995. These comparisons will be conducted for water column TSS (or other sorbent compartments) and PCBs, sediment PCBs, and PCBs in fish.

Because the time scale over which the models are to be applied spans decades, some aggregation of inputs with respect to fine time-scale detail is appropriate. The goodness of fit of that short-term simulation should be evaluated in accordance with the temporal scale to which the model is intended to be applied.

Based on suggested evaluation methods by Reckow et al., 1988 and Thomann, 1982, the following metrics and quality criteria are specified:

a) Graphical methods. This will involve graphical comparisons of model predictions versus observed data, in order to qualitatively assess the goodness of fit of calibration. It will serve to complement the quantitative statistical evaluations discussed below by highlighting any strong relationships between prediction errors and key spatial, temporal, or causal elements of the system. Specific analyses include: time series, point-in-time, and temporal and spatial distributions. Presuming data are available to make a comparison, specific locations to apply this metric include: Appleton, Kaukauna, Little Rapids, DePere, the Fox River mouth, and Green Bay.

b) Comparisons of distributions of predicted vs. observed values. This metric will compare temporal and spatial distributions of predicted and observed values for: water-column TSS and PCB concentrations, and of PCB body burdens in fish, distributed over time and within river reaches and portions of Green Bay. In an ideal situation, the mean values and slopes of the predicted and observed distributions would be identical. However, given the uncertainties in both the forcing functions and parameter values that control model predictions as well as uncertainty in observations, it is expected that the average values predicted and observed should agree to within 30%. The nonparametric Kolomogorov-Smirnov test will be used to compare distributions, using 0.05 and 0.10 confidence levels (for a one-tailed test) to evaluate whether the predicted and

observed distributions are similar.

c) Event and non-event concentration and flux comparisons. Advective transport of particles, particle-phase PCBs, as well as dissolved PCBs is an important mechanism that affects the fate of PCBs in the Fox River/Green Bay system. Because there are sediment resuspension processes that occur during periods when high shear stresses are exerted on the sediment-bed (typically high flow periods in the river or wind and seiche events in the bay), and because PCBs are particle-associated, contaminant transport may be greater during events than during non-event periods.

This metric compares predicted versus concentrations and, where possible, fluxes of TSS and PCB for periods when shear stresses at the sediment water interface are estimated to exceed the critical shear stress at which quantitative resuspension of the sediment bed occurs. Application of this metric will be limited by the availability of data for both event and non-event conditions. To properly apply this metric it will be necessary to: 1) identify physical conditions (flow, wind speed and direction, circulation pattern, etc.) that delineate event and non-event conditions; 2) identify an effective representation of these physical conditions in the models (e.g. river flow rate or bay wind speed at which the average shear stresses at the sediment water interface exceeds some limit); 3) define limits to properly aggregate model predictions and observations for comparison (e.g. for a point-in time spatial composite or a point-in-space temporal composite); and 4) define the quality criteria for this analysis. Presuming data are available to make such a comparison, specific locations to apply this metric include: Appleton, Kaukauna, Little Rapids, DePere, the Fox River mouth, and Green Bay.

Model predicted fluxes will be compared to independent estimates of fluxes using the statistical methods evaluated by Preston et al. (1989) and available measurements of flow and water column concentrations. These comparisons between model predicted and statistically estimated fluxes are intended to provide additional checks of the goodness-of-fit of model calibration. It is important to understand the potential limitations of flux comparisons as a metric. Fluxes computed using statistical estimation methods are themselves model predictions, in this case a statistical model. The true magnitude of fluxes cannot be established because measurements are not available at sufficiently high temporal resolution. As a result, only a relative comparison between these two types of fluxes is possible. Differences between the mass balance and statistical flux estimates are not in themselves indications of prediction biases in either method.

In an ideal situation, the mean values and slopes of the predicted and observed concentrations and estimated fluxes would be identical. However, given the uncertainties in both the forcing functions and parameter values that control model predictions as well as uncertainty in observations, it is expected that the average values predicted and observed should agree to within 30%.

d) End-of-period mass budget analysis. This metric will compare predicted and observed PCB mass distribution, by compartment. Comparisons will be made on an order-of-magnitude basis; however, predicted and observed PCB mass inventories should agree to within 30%. The 1989 Green Bay Mass Balance Study data define the initial sediment PCB distributions/masses for the river and bay models. The 1992-1995 remedial investigations define end-of-period conditions for select locations (Deposit POG, N, etc.) in the river upstream of the DePere Dam. The 1995 remedial investigation defines end-of-period conditions for the river downstream of the DePere

Dam. Data collected for Green Bay since the 1989 study, if any are available, will be used to define end-of-period conditions for the bay to conduct this analysis for the bay model.

e) Sediment Bed elevation changes. This metric compares predicted changes in sediment bed elevation with monitored changes in bed elevations as determined from bed elevation transect data as well as dredging records. Sediment bed elevation exists for several locations in the Fox River upstream as well as downstream of the DePere Dam; data may also exist, in the form of dredging records, for the navigation channel maintained in Green Bay. To perform this analysis, it will be necessary to rectify each set of sediment bed elevation observations to a single, consistent elevation datum. Comparisons will be on a trend and magnitude basis; however, average predicted and observed bed elevation changes should agree to within 30%.

f) Net burial rates. This metric compares predicted net burial rates to rates estimated from Cesium-dated sediment core data for the Lower Fox River and Green Bay. Comparisons will be on a trend and magnitude basis. However, for locations where the mechanisms of sediment accumulation can be distinguished (deposition versus slumping of a dredged channel) and where conditions at core site are representative of conditions over a larger area (such as a the area represented by a model sediment segment) average predicted and estimated rates should agree to within 30%.

It is important to recognize the limitations of this metric, especially as applied to dynamic systems such as rivers. First, the source of Cesium to a waterbody is through direct deposition to the water surface or the deposition to the watershed with subsequent runoff to the waterbody. For the Lower Fox River, the surface area of the watershed is very large relative to the surface area of the water. Because particle delivery from the watershed significantly varies from year to year (expressed by the variability of rainfall-runoff events), the year in which particles were deposited to the watershed surface and the year in which those Cesium-laden particles become incorporated into the sediment bed can differ significantly. As a result, the Cesium profile at a location the river may be distorted in some unquantifiable manner. Second, erosion and deposition events may further distort any Cesium signal still remaining after the radionuclide is first delivered to the sediment bed. These factors can also influence the quality of Cesium profiles in larger waterbodies, especially in nearshore areas influenced by runoff, erosion, and deposition events.

g) Uncertainty analysis. This metric characterizes the precision of model predictions. Model uncertainty has two sources: 1) uncertainty in information used to define model forcing functions (forcing function uncertainty); and 2) uncertainty in the values assigned to model parameters for a given environmental condition (parameter uncertainty). The key feature of typical short-term, data rich periods is that the information used to define model forcing functions is relatively defined; as a result, forcing function uncertainty is low. Given the high quality, comprehensive data sets available to characterize short-term forcing functions for the Fox River and Green Bay models, it is assumed that all model uncertainty is attributable to parameter uncertainty for short-term simulations. The methodology to perform the uncertainty analysis is still under discussion by the Model Evaluation Workgroup.

3.3 HINDCAST METRICS

A historical hindcast beginning in 1957 and ending in 1995 will be a valuable tool to evaluate

model performance over long time frames. In 1957, the Fox River and Green Bay are believed to have been essentially unimpacted by PCBs. This is represented by assigning zero as the initial PCB concentrations in all water, sediment, and biota compartments of all models. The hindcast is then conducted by applying the same simulation procedures used for the short-term simulation to assign values for each model parameter, and completing a simulation that starts from this initially unimpacted state and continues through contemporary period(s) of high data availability and certainty. Model predictions are compared to observations for these windows of high data availability. The more closely predictions and observations agree the stronger the predictive ability of the procedures used to assign model parameters for conditions outside of the short-term calibration. This establishes the predictive capabilities of the models.

Two categories of hindcast metrics will be employed: 1) the goodness-of-fit to observations, and 2) uncertainty analysis. The goodness-of-fit metrics are the same metrics as those used to evaluate the short-term simulations and include: a) graphical methods, b) comparisons of predicted and observed distributions, c) event and non-event comparisons, d) end-of-period mass budget analysis, e) sediment bed elevation changes, and f) net burial rates. The methodology to perform the uncertainty analysis is still under discussion by the Model Evaluation Workgroup.

It is important to recognize that data gaps exist throughout the hindcast period. In particular, loading information for PCBs and solids is more limited and uncertain for the period 1957-1989 than it is for the period 1989-1995. The physical characteristics of solids discharged from point sources are also known to have been significantly different early in the hindcast period than in more recent years. As a result, model forcing functions estimated from these data, such as loads, will be far more uncertain than those estimates for periods of high data availability.

It is also important to distinguish between the uncertainty attributable to the limited nature of information used to develop forcing functions and the uncertainty attributable to model process parameterization. This is necessary to prevent shifting the focus of the hindcast analysis from model evaluation to an unconstrained exercise in model parameter calibration. Prediction uncertainty that is a consequence of uncertainty in model forcing functions is not a short-coming of model parameterization; parameterization error is not indicated by the hindcast analysis unless observations fall outside the uncertainty bounds of hindcast predictions (i.e. the prediction uncertainty attributable to uncertainty model forcing functions).

As a result, the uncertainty analysis of the hindcast simulation differs from that for the short-term simulation. For the hindcast, all model uncertainty will be assumed to be attributable to forcing function uncertainty. If observations fall within the uncertainty bounds of hindcast predictions, then all model error can potentially be described by uncertainty in model forcing functions and the model parameterization is judged acceptable (i.e. there is no need to revise the methods used to assign model parameter values). If observations do not fall within the uncertainty bounds of the predictions, then forcing function uncertainty cannot account for all model error and the model parameterization is judged inadequate (i.e. the methods used to assign model parameter values must be refined).

3.4 FORECAST METRICS

Forecast metrics are used for two purposes. The first use is as a comparative tool to differentiate the quality of competing model alternatives that have successfully met the quality criteria for all preceding metrics. In this use, the model alternative with the smallest uncertainty bounds of predictions is the preferred model alternative. The second use is as a tool to compare predicted outcomes of environmental management strategies for the preferred suite of models selected as a result of the model evaluation process (i.e. compare the series of forecast simulations for no action and other alternative scenarios). In this use, uncertainty bounds are used to determine whether the alternative scenarios explored generate predictions that are different, in the sense that their uncertainty bounds have nonoverlapping portions.

Uncertainty bounds of predictions. To apply this metric, all model uncertainty is assumed to be attributable to model parameterization error (i.e. future loads and other forcing functions are known). An uncertainty envelope must be generated for each model parameter that contributes to prediction uncertainty. In the Fox River and southern third of Green Bay, the long-term fate of PCBs is believed to be controlled by the dynamics of the particles with which PCBs associate. Therefore the key uncertain parameters are settling velocities, resuspension velocities, sorbent transformation/exchange rates. The uncertainty envelope (mean, minimum, and maximum values, coefficient of variation, etc.) for each of these parameters is determined at the stage of model calibration and short-term simulation development (i.e. the data used to develop procedures to assign model parameters indicate the range of input uncertainty). If available data for the Fox River and Green Bay are not sufficient to define parameter uncertainty envelopes, such as may be the case for sorbent transformation/exchange rates, theoretical knowledge or data for other physically similar natural systems may be used to further support development of uncertainty envelopes. The same time series of environmental input variables (hydrograph, loads, settling and resuspension velocities, etc.) are employed in each set of simulations.

It is again important to distinguish between uncertainty that is attributable to uncertainty in forcing functions and the uncertainty attributable to model process parameterization. The ability to address the former can be very limited, and the resulting uncertainty is not a short-coming of the model itself. It is for this reason that the forecast uncertainty metrics emphasize uncertainty attributable to process parameterization.

4.0 PRIORITIZATION OF METRICS

The model evaluation metrics described in the preceding sections are intended to be applied in the systematic and sequential manner identified below:

1. Mathematical Representation of the Natural System

   The mathematical representation of the physical, chemical, and biological characteristics of the modeled system should be consistent with the observed behavior of the real system and appropriate for the planned application, including spatial, temporal, and kinetic contexts. Each environmental process/pathway that can significantly affect contaminant distribution and long-term fate must be included in the mathematical representation of the natural system.

2. Short-Term and Long-Term Simulation Metrics

   Short-term simulations are used to: 1) determine appropriate ranges of values for each model parameter based on analysis of data for time periods where data quality/certainty is greatest; and 2) develop generalized methods to assign model parameter values for any given environmental conditions as a function of independent observations (flow, wind speed and direction, temperature, etc.). Model performance is characterized by graphical methods, distribution comparisons, event and non-event concentration/flux comparisons, mass budget analyses, sediment bed elevation comparisons, net burial rate comparisons, and uncertainty analysis. These metrics establish the character (accuracy and precision) of short-term model performance.

   Long-term, retrospective hindcast simulations are used to: 1) provide a check of the methods used to assign model parameters over long timeframes as well as conditions outside of the short-term calibration; and 2) establish the predictive capabilities of a model by comparative analysis of hindcast results to short-term simulation results for the same period. Model performance is characterized by the same metrics as are used for short-term simulation and include: graphical methods, distribution comparisons, event and non-event concentration comparisons, mass budget analyses, sediment bed elevation comparisons, net burial rate comparisons, and uncertainty analysis. These metrics establish the general character of model performance (accuracy and precision) as a tool for conducting long-term, predictive simulations.

3. Forecast Simulation Metrics

   Forecast simulation metrics are used to: 1) provide a final metric to differentiate the quality of competing model alternatives that have successfully met the quality criteria for all preceding metrics; and 2) to determine whether alternative scenarios explored using the selected suite of models generate predictions that are different, in the sense that their uncertainty bounds have significant nonoverlapping portions. Model performance is characterized by uncertainty analysis. This metric establishes the precision of model predictions.

5.0 REFERENCES

Reckow, Kenneth H., et al., "Statistical Evaluation of Mechanistic Water-Quality Models", Journal of Environmental Engineering, 116:2, March/April 1990.

Thomann, Robert V., "Verification of Water Quality Models", Journal of the Environmental Engineering Division, Proceedings of the ASCE, 108:EE5 , October, 1982.

Preston, S. D., V. J. Bierman, Jr., and S. E. Silliman, "An Evaluation of Methods for the Estimation of Tributary Mass Loads", Water Resources Research, 25(6), 1989.